DOCUMENT RESUME

ED 319 784 TM 015 018

AUTHOR L'Hommedieu, Randi; And Others
TITLE Putting the "But" Back in Meta-Analysis: Issues Affecting the Validity of Quantitative Reviews.
PUB DATE Apr 87
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Analysis; Literature Reviews; *Meta Analysis; *Research Methodology; Research Problems; Research Reports; *Statistical Analysis; *Validity

ABSTRACT

Some of the frustrations inherent in trying to incorporate qualifications of statistical results into meta-analysis are reviewed, and some solutions are proposed to prevent the loss of information in meta-analytic reports. The validity of a meta-analysis depends on several factors, including the: thoroughness of the literature search; selection of studies for inclusion; appropriate coding and analysis of studies; and report format selected. The solution proposed to the problem of methodological quality is to include all selected studies and report an average effect size for the aggregate. The report on the meta-analysis then should be a qualitative, discursive argument rather than a simple statistic. Proposals for putting the "but" back in meta-analysis are: (1) assure that it is not a substitute for qualitative review; (2) offer the reader information necessary to evaluate the validity of decisions made at the individual level; and (3) assure that qualifications of studies are not excluded from the analysis. A thorough quantitative review should include: a discursive review of each study; a report on how each effect size was calculated; the location of the summary statistics upon which each effect size was based; and a discussion of study limitations and the factors that affect validity of effect size. Suggestions are also given for appropriate reporting and avoiding publication bias. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED319784

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
RANDI L'HOMMEDIEU
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Putting the "But" Back in Meta-Analysis: Issues Affecting the Validity of Quantitative Reviews

Randi L'Hommedieu, School of Music
Robert J. Menges and Kathleen T. Brinko, Center for the Teaching Professions

Northwestern University,
Evanston, Illinois

April 12, 1987

BEST COPY AVAILABLE

Prepared for the meeting of the American Educational Research Association,
Washington, DC
April, 1987

TM015018

Last year we undertook a meta-analysis of studies involving student ratings feedback to college instructors. We began with a basic knowledge of meta-analysis procedures, a strong background in the feedback literature, and a previous meta-analysis on the same topic (Cohen, 1980) to serve as a model.

During the course of the project we encountered a host of problems and made numerous decisions, each of which influenced our final results. As these problems multiplied, we found it increasingly necessary to qualify our statistical results. We became both frustrated at the difficulty of incorporating these qualifications into the meta-analysis, and alarmed at the amount of information that may be lost in typical meta-analysis reports. We began to search for ways to put the "but" back in meta-analysis. This paper documents some of our frustrations and describes some of our proposed solutions.

The validity issues that we faced fit into three broad categories. The first deals with issues involving the calculation of effect sizes and reporting of the meta-analysis. The second validity issue concerns the quality of the research that we are integrating. The third issue has to do with the professional context in which research is conducted and published.

Meta-Analysis Methodology and Reporting

We undertook our project with the following plan. We would thoroughly search the literature for relevant research conducted since Cohen's 1980 meta-analysis. We would code all of the studies, using 36 variables that we thought might have a bearing on the findings. We would then calculate a magnitude of effect standardized against the standard deviation of the control group for each study in the analysis. Finally, we would compute an average effect size for all studies and discuss the mediating effects of the 36 study characteristics.

The meta-analytic literature had prepared us for some of the problems we would face. We were not prepared, however, for the frequency with which this straightforward plan was frustrated, requiring us to choose from among methodological alternatives that had a significant impact on our results. As the project progressed, our objective, mechanical integration of the literature seemed to be evolving into a morass of informed, but arguable decisions and compromises. Worst of all, we realized that our intended report format would document only the grossest of these decisions. Our analysis would not, in any practical sense, be replicable. We realized, however, that by the standards of many journals it would nevertheless be publishable.

Search of the Literature

The validity of a meta-analysis depends upon a complete, or at least representative, survey of the population of interest. Unfortunately, the results or study characteristics of individual studies are often related to how easily the study comes up in a casual search. Publication bias is the most often cited example. However, it is also easy to overlook dissertations written before establishment of the DAI database, internal reports of faculty development programs, unpublished research reports, research from other fields, masters's theses, and reports made available after the original search was terminated.

The methodological literature has not underplayed the importance of these procedures and we agree that no decisions or compromises are justifiable at this stage of a meta-analysis. However, we can report that a continuing, exhaustive search of the literature is tedious and tempts the reviewer to pursue only the most promising leads. Following the first report of our preliminary results at last year's meeting, we uncovered additional studies which required slight revisions of our findings. Indeed, other sessions at this year's meeting hold papers that qualify for inclusion. Since meta-analyses do not completely resolve issues and prevent additional research, it is important to view the meta-analysis report as a static picture of a dynamic process. We think that the best meta-analyses will be set up as ongoing projects with periodic reports.

Selection of Studies for Inclusion

Selection criteria are initially determined by the research question and the parameters set by the meta-analyst. These early decisions have a profound effect on the final result, yet they may be based on methodological as well as substantive reasons. Reviewers working with the same population of studies and with the same research questions can arrive at different conclusions simply because their selection criteria pulled different studies from the literature.

We have re-examined our selection criteria several times over the course of the last year. Each re-examination required another review of the previously rejected studies. And we found that the casual examination for inclusion that we had anticipated was seldom adequate. What had appeared to be a fairly straightforward, objective process was frustrated by two factors: (1) our own changes of mind as we became increasing familiar with the literature, and, relatedly, (2) the needlessly arcane and obfuscatory reporting style that was a feature of many studies.

Our final decision regarding selection was to include any study in which some measure of the effect of student ratings feedback could be calculated or reliably estimated. However, with this broad criterion for inclusion, we considered it important that the final report should provide adequate information for readers who disagree with our selection procedures to re-do the analysis using their own selection criteria.

Coding and Analysis of Studies

The coding and analysis of individual studies presented us with our most bewildering array of choices. In coding the studies we often found it necessary to interpolate, estimate, or even code as missing some variables that were of interest to us, but were not important to the author of the study.

In computing effect sizes we ran into the expected problems of converting nonparametric statistics and estimating effect sizes from T and F values. In addition, our computations brought out two problems that, while quite obvious, had not occurred to us before undertaking the analysis. First, use of the T- or F-value effect size formula will generally yield a more conservative effect size than the Z-score formula. This is because the T- and F- formulas are based on a pooled estimate of group variation. Whether or not other researchers choose, for the sake of equivalency, to base all of their effect size measures on pooled estimates of variance, meta-analysts and consumers of meta-analyses should be aware of how this factor effects the comparability of measures.

A second unforseen problem with the use of F-values is that F gives no indication of the sign of the effect. In several studies in which the treatment did not show statistically significant differences, the author provided F-tables, but did not provide group means. It is possible to compute an effect size in such cases, but it is impossible to tell whether the effect favors the treatment or the control group. In one exceptional case, the author failed to report the sign for a statistically significant effect!

Another issue regarding the comparability of effect sizes has to do with the comparison of adjusted and unadjusted criterion scores. Although whether to adjust scores based on initial differences or to rely on random assignment alone to equate groups is a decision for the individual researcher, the sole reason for such adjustment is to arrive at a larger F-value. We believe that this is another issue to be considered when comparing effect sizes across studies.

Our most uncomfortable analysis decisions were caused when we could not compute a reliable effect size from the information provided in the report. In all cases we tried to locate more complete reports (dissertations or ERIC documents), recalculate from the raw data, make informed estimates based on related information, and contact the authors for additional information. However, in eight of the thirty studies in our meta-analysis these efforts still did not yield a reliable effect size. In five of the eight cases the group differences seemed so small and random that we were confident in assigning an effect size of zero to the study. In the three remaining studies, we opted to assign an effect size equal to the average of similar studies with comparable results. This was hardly a satisfactory solution, but it was the best that was available.

A further methodological issue deals with the number of effect sizes computed for each

study. Multiple comparisons introduce problems with dependence of measures and the weighting of individual effect sizes. The use of cognitive and affective dependent measures in student ratings research makes this issue even more complex.

Our choice was to avoid these complications by computing separate effect sizes for three types of dependent variable: student ratings, affective outcomes, and cognitive outcomes. Within each study we averaged the effect sizes of multiple comparisons to obtain a single effect size for each study. But we retained the effect sizes for each subcomparison for use in eventual subanalyses of results.

Even with this simple plan, subjective decisions became necessary that influenced our interpretation of individual studies. In Hoyt and Howard (1978), for instance, we were able to calculate three different effect sizes ranging from 1.10 to .778. Each of the alternatives could be justified, yet they resulted in significantly different figures to be contributed to the analysis.

Meta-Analysis Report Format

Meta analysis has not developed a standardized reporting format as have other research procedures. However, the typical meta-analysis report includes (1) a detailed description of the search and selection criteria, (2) a description of the coding procedures, and (3) a cross-tabular presentation of the coded information and effect sizes.

It became clear early on in the project that a report of this type would mask important decisions that readers would need to consider when evaluating our findings. We were determined to arrive at a report style that would give the reader the opportunity to evaluate the logic of the decisions we made and, more importantly, to trace those decisions back to the analysis. Readers should be able to re-do our calculations using the information provided in our report. And so we determined that in addition to the traditional meta-analysis features listed above, our report should include (1) an explanation of our methods for computing each effect size, (2) the location of the statistics that we used for this calculation, and (3) a discussion of all the features that affected our confidence in the computed effect size. These considerations led us to the conclusion that the quantitative synthesis of study results alone would fail to adequately convey the information that our exploration of the literature provided. It became clear to us that the meta-analysis could not be a substitute for a discursive, qualitative review of the literature, but rather should be an extension of the traditional review.

Quality of the Research Base

The studies included in our review of the feedback literature vary widely in methodological quality. Our survey of the literature ranged from studies that we would propose as exemplars of inquiry in our field to very weak studies in which data were discarded, analyses were confused, procedures were bungled, and reports were misleading. Yet, even the worst reports contained information that could be of use--if only in a negative way.

As researchers, as well as reviewers of research, we are particularly sensitive to the tendency of reviewers to become "Monday morning quarterbacks." Field research in the social sciences, particularly in education, is conducted under a set of practical, ethical, and economic restrictions that makes methodological compromise a fact of life. Furthermore, we agree with methodologists like Cronbach and Dunn who complain that blind attention to methodological purity can yield studies with indisputable, but trivial and useless, results.

For all of these reasons and more, we were reluctant to reject studies for purely methodological reasons. To account for the variable quality of studies we were synthesizing, we had originally decided to code studies according to established elements of good methodology. However, we soon ran into several instances where this procedure broke down. For example, most of the studies in our review used random assignment to treatment, yet in almost all of these cases attrition compromised the original equivalency of the groups. Were we to rate these studies more favorably than a quasi-experiment in which initial differences were statistically controlled? Should we rate a study that used only selected items from a standardized ratings instrument more favorably than a study that used a carefully constructed, but nonstandard instrument? Each of these individual questions is answerable. But a final, summative rating that takes all of these questions into account obscures the complexity of these important issues. We eventually despaired of devising a "rating" scale that would accurately reflect the delicate balance between the procedures of a study and the value of the final results.

Our solution to the issue of methodological quality is to include all selected studies and report an average effect size for the aggregate. In addition, in our subcomparisons we will report an average effect size for those studies that we believe show the most rigorous and valid tests of the effects of feedback. Our report will contain our justification for selection of these studies, but it will be a qualitative, discursive argument rather than a simple statistic. In addition, our discursive review will contain all the information necessary for critical readers to challenge our proposal and to restructure the analysis to meet their own criteria.

The Social Context of Research

The bulk of research in education is "required" research. Dissertation projects are the most obvious example, but no less required is the research that many college faculty produce for tenure or advancement. Reviewers of research in our field must remember that the quality of research is often limited by inadequate funding, minimal institutional support, and alternative demands for the researcher's time and attention.

While we see no advantage to gnashing our teeth over this research tradition, we think it is vitally important that researchers and journal referees and editors take their responsibilities more seriously. We have read published reports of projects in which unwanted data were purposefully discarded and studies in which only the results of the statistically significant tests were reported. Compounding these methodological problems was the reporting style. We often found ourselves extremely frustrated by reports requiring several hours and days of re-reading and detective-style cross-referencing of documents. We believe that even fairly thorough reviewers may have seriously misinterpreted the results of some of these studies. Perhaps one of the advantages of a meta-analysis requiring a substantive, interpretable, and replicable effect size is that it forces out problems that are not revealed by less thorough review methods.

Much has been written about "publication bias," that is, the tendency for journals to favor for publication studies that result in statistically significant findings. Reviewers have expressed concern that this tendency has resulted in a published literature that contains an inflated proportion of Type I errors. This makes it particularly important that meta-analysts search the unpublished literature and the unsubmitted "file-drawer" literature in order to reach an unbiased measure of effect.

We have hinted at an even more insidious result of publication bias. We are concerned with the possibility that reports are being written and research is being presented in a manner that emphasizes significant results in order to enhance its potential for publication. We hope that the field's research practitioners can agree to police themselves and that editors can insist on a reporting style that is clear, complete, concise, accurate, and objective.

Conclusions and Recommendations

Colleagues have asked us why we are attempting to integrate a literature of which we are so critical with a technique we distrust. Our answer is that a "traditional" meta-analysis--a summary description of aggregates conducted with the intention of resolving a research issue--is clearly out of the question. Cohen's meta-analysis has not discouraged additional research. Indeed, we were

surprised to find that this important work is usually cited with no more authority than an individual study. We want to continue with the statistical integration of the quantitative literature on ratings feedback, but we need a way to introduce the myriad qualifications that we feel are essential for the intelligent interpretation of results. We want to put the "but" back in meta-analysis. Here is what we propose.

"But" No. 1

Meta-analysis is an important addition to what may become a rigorous review methodology, but it is not a substitute for the qualitative review. Meta-analysis was proposed as an answer to the problem of unmanageable literatures. It was to be a means to avoid synoptic, qualitative analyses. Yet, the encyclopedic review of literature is still important. Statistical integration is a descriptive, summary tool that should be used in conjunction with, not instead of, the qualitative review. Perhaps more intelligent review and synthesis of the literature will educate us to redirect our research resources so that unmanageable literatures do not accumulate.

Of course, the qualitative review is not above criticism. We object to the qualitative review that simply rewords the author's summary section, just as we object to the meta-analysis that skims the document for coding variables and calculates an effect size from misunderstood statistics. There is a tradition of criticizing expediencies in single study research, but we haven't so far applied the same standards to literature reviews.

"But" No. 2

An effect size is a useful tool for producing standardized measures of effect, but, like any descriptive statistic, its summative nature masks important subjective decisions at the individual level. We thought that the analysis and coding of studies would be a simple and objective procedure, however in well over half of the studies in our analysis we had to choose from among several alternative methods to calculate or estimate an effect size. We are confident that our calculations provide useful information about the individual studies and some notion about the general tendencies of this body of literature, but our methods are not beyond criticism. Meta-analyses should offer the reader the necessary information to evaluate the validity of these decisions.

"But" No. 3

Meta-analysis offers a way to obtain comparable measures across studies with different questions, methodologies, and procedures, but it does not diminish the importance of these

differences to the interpretation of results. Hunter, Schmidt and Jackson (1980) answer the charge that meta-analysis "combines apples and oranges" with the idea that meta-analysis is useful for studying "fruit." We agree, but these qualifications need not and should not be excluded from the analysis. The coding of variables and the regressions against study characteristics are helpful techniques, but not sufficient.

Recommendations

The meta-analysis report format has an authoritative, objective veneer, but this veneer masks equivocal methods and findings, prevents re-analysis, and discourages further inquiry. We suggest that meta-analyses be considered only a part of a comprehensive review. In addition to the standard information included in both qualitative reviews and meta-analyses, a thorough quantitative review should include:

1. A discursive review of each study;
2. A report of how each effect size was calculated;
3. The location of the summary statistics on which each effect size was based;
4. A discussion of the study limitations and the factors that affect the validity of the effect size.

If the review is thorough, the reader should be able to trace the review findings back to the literature. Single study authors operate under ethical constraints that make blind aggregation necessary. Reviewers are under no such obligation. Indeed, providing information that leads the reader back to the results of the individual study is precisely what is required.

We also believe that this type of research review has implications for authors of individual studies. Based on our experience with this literature, we offer some general suggestions for the authors of single studies that will facilitate their inclusion in future meta-analyses and will coincidentally improve the methodology and reporting of the individual study.

Report substantive measures of results--group n's, means, standard deviations, etc.--for every comparison and test. We understand that, empirically, a statistical test is based on the idea that the observed mean is an imprecise and variable estimation of the population mean. But for substantive reasons--even if only for the sake of completeness--the mean and SD of each comparison is essential. We were appalled to discover how many studies neglected to include this simple information.

Report all results, regardless of statistical significance. Even nonsignificant results hold interest for some readers--particularly meta-analysts. Furthermore, we believe that reporting and

discussing only significant differences is a less than honest way to present the results of a study.

Pay particular attention to the validity of the construct to be operationalized. The research question that we wished to explore was this: Do student ratings, as typically administered at colleges and universities, stimulate instructional correctives that result in higher subsequent ratings? But several studies that claimed to share this purpose were conducted at institutions that had standing student evaluation procedures. In student rating research, the midterm rating intervention is an experimental artifact made necessary because of the desirability of using the same students and teachers on the pre and post measures. We should not confuse this operationalization with the construct of interest. It seems important, therefore, that subjects in ratings feedback studies be teachers who have had no previous--or at least no recent--experience with ratings scales.

Strive for a reporting style that is clear and concise. While we realize that research requires very precise vocabulary, we often found the language to be needlessly arcane and stilted. Negative examples would be illustrative, but unkind. We can, however, cite McLean (1979) as one example of research prose that is clear and succinct, yet in no way compromises the rigor of the report.

Journal editors, referees and dissertation committees need to share in these responsibilities. Editors should also consider the effects of publication bias. We believe that studies that are well designed and analyzed, that explore significant questions thoughtfully, and are reported well have always been publishable, regardless of results. Critics of publication bias often tacitly imply that there are enough studies awaiting publication that fine studies that fail to show significant differences are being excluded from the literature. This may not be true. More selective publication of research will make the published literature more valid and representative, but it will consequently increase the "file drawer" problem. Institutions can help by providing more support for serious researchers and by recognizing the other important contributions of faculty who are not.

Meta-analysis is a methodological refinement based on the same logic as the single study and it shares all of the inferential limitations of the single study paradigm. The problems of the single study are not solved by meta-analysis, rather, they become "meta-problems." We believe that the idea that meta-analysis can completely "resolve" a research question is only wishful thinking.

Our meta-analysis of student ratings feedback research has shaken our faith in the power of this technique to reach indisputable conclusions, but we have a new appreciation of how meta-analysis can organize and inform an ongoing research program. With more modest aspirations and cautious application it can supplement other synthesis procedures and contribute to more rigorous research in many areas of inquiry.

REFERENCES

Bracey, G. (1987). The time has come to abolish research journals: too many are writing too much about too little. Chronicle of Higher Education.

Centra, J. (1973). Effectiveness of student feedback in modifying college instruction. Journal of Educational Psychology, 655, 395-401.

Cohen, P. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. Research in Higher Education, 13, 321-41.

Cook, T.; & Leviton, L. (1980) Meta-analysis critically assessed, Journal of Personality, 1980.

Cronbach, L. (1983). Designing Evaluations of Educational and Social Programs. San Francisco: Jossey-Bass.

Dunn, W. (1982). Reforms as Arguments. in Knowledge: Creation, Diffusion, Utilization,3, 293-326.

Glass, G.; McGaw, B.; & Smith, M. (1981). Meta-Analysis in Social Research. Beverly Hills, CA: Sage.

Hoyt, D.; & Howard, G. (1978). The evaluation of faculty development programs. Research in Higher Education, 8, 25-38.

Hunter, J.; Schmidt, F.; & Jackson, G. (1982). Meta-Analysis: Cumulating Research Findings Across Studies. Beverly Hills, CA: Sage.

Kuhn, T. (1963). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

McLean, D. (1979). The effects of midsemester feedback upon weekly evaluations of university instructors. Unpublished master's thesis, Department of Psychology, University of Western Ontario, London, Ontario.

Menges, R. & Brinko, K. (1986) Effects of student Evaluation Feedback: a meta-analysis of higher education research. Evanston, IL: Center for the Teaching Professions, Northwestern University.

Rotem, A. (1978). The effects of feedback from students to university professors: an experimental study. Research in Higher Education, 9, 303-18.